

# Weekly Report

07/12/2015-07/19/2015

## Research

### 统计在家人数

本周首先根据基站数据统计了每一个时段（按一小时划分）在家的人数和总人数。在家人数按照1小时内和家的平均距离来判断，如果离家距离小于500米，则认为在家。总人数是这个时间段内统计到的所有手机用户。图是第一次统计出来的结果，发现在凌晨4到6点间的在家人数最少（更正前的数据），但这和常识是相矛盾的。

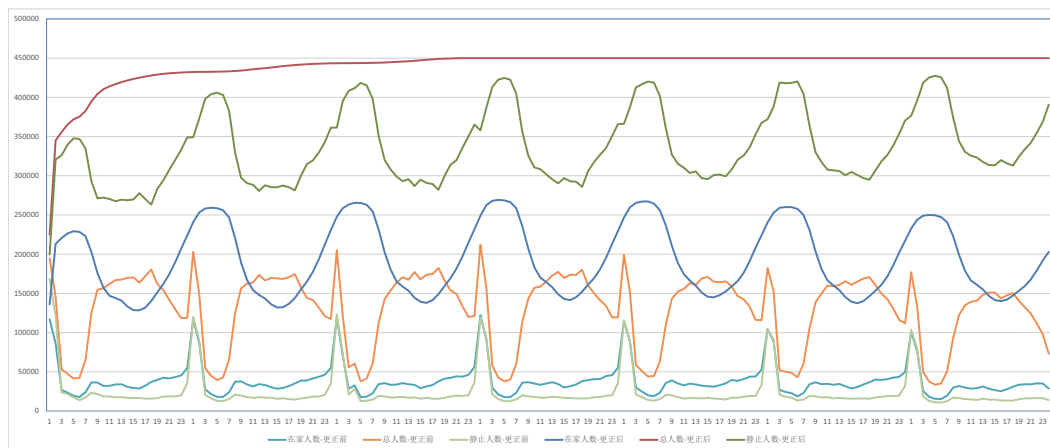


Figure 1: 人数统计

同时，我们看到在4到6点间的总人数也降低了，这是因为是由于手机用户静止或者关机导致在这个时间段内没有信号纪录，于是我们的总人数减少了很多，相应的在这个点的在家人数就少了。图1显示的是我们对轨迹分割采用的分割方法，在时

间轴上橙色点是原始的数据，黄色点是我们人为添加的数据。比如切割cd时间段内的轨迹时，因为一头一尾没有数据点，我们都向前取一个数据点，保持基站号相同，时间在相对应的c点和d点上。然而我们之前的代码漏考虑了一种情况，导致在分割de段的时候没有添加点，程序认为这段没有数据点就不记录，以至于人数不累加进去。于是在我们对这个问题更正后，从图中我们首先可以看到，总人数

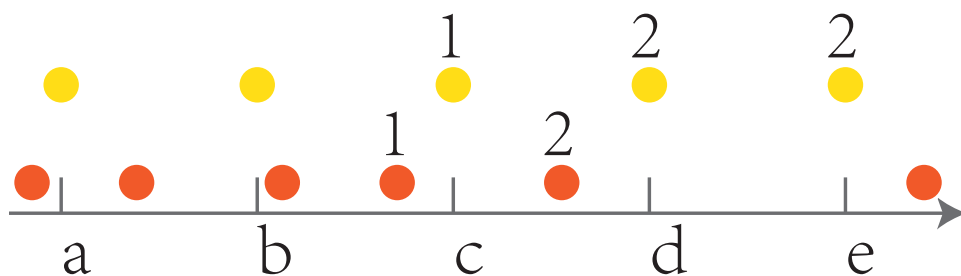


Figure 2: 轨迹分割

一直在上升直到平衡，这是因为这个人一旦出现就会记录下来，并且在之后的轨迹分割中一定还会出现。在家人数的数量有着很大的提升，并且较更正前更加平滑，在家人数最低点出现在下午3点左右。

## 为kmeans寻找最好的K

我把1万人的数据（25万个10维数据点）使用Matlab的内置函数进行对kmeans聚类质量的评估。

```
%使用CalinskiHarabasz
k=[2 3 4 5 6 7 8 9 10 20 30 40 50 60 70 80 90 100];
clust = zeros(size(X,1),size(k,2));
opts = statset('MaxIter',1000);
for i=1:size(k,2)
    clust(:,i) = kmeans(X,k(i),'Options',opts);
end
eva = evalclusters(X,clust,'CalinskiHarabasz')

%使用silhouette
opts = statset('MaxIter',1000);
myfunckmeans = @(X,K)(kmeans(X, K, 'Options',opts));
eva = evalclusters(X,myfunckmeans,'silhouette','KList',
    ,[1:10 15:5:100])
```

因为silhouette的复杂度太高，计算时间太长了，最后在计算25万个数据点时放弃silhouette，转而使用CalinskiHarabasz。结果如表1所示，最好的k是2，在2万个点的数据集上使用silhouette的结果也是2，所以对于我们抽取出来的特征在k的选择上使用CalinskiHarabasz等评判标准不太适用。

k	2	3	4	5	6	7	8	9	10
CriterionValues	139,968	129,254	113,726	100,055	91,379	91,011	84,869	82,893	80,166
k	20	30	40	50	60	70	80	90	100
CriterionValues	57,117	46,917	40,472	36,303	32,844	30,308	27,797	26,407	25,010

Table 1: CalinskiHarabasz

## Plan for next week

- 开始编程训练。
- 代码更正后，静止人数有了巨大的提升，而这些提升我认为这是由于处理了太多de段的数据造成的，下周研究一下这一问题，提取轨迹信息更丰富的人可能可以解决这个问题。